5/2025, sentyabr-oktyabr (№ 00079)

DATA SCIENCE-DRIVEN APPROACHES TO IDENTIFYING AI-GENERATED CONTENT: MACHINE LEARNING AND NLP MODELS FOR ACADEMIC INTEGRITY AND DIGITAL TRANSPARENCY

Abdullaev Munis Kurbonovich
Head of the Department of "Industrial Management and Digital Technologies" of the International Nordic University, PhD, Professor. Independent Researcher of TSUE
Email: m.abdullayev@nordicuniversity.org
ORCID: https://orcid.org/0000-0003-0290-8453

Kungratov Ilmurod Kuzibay ugli
Master's student (data science) in International Nordic University. Scientific journals editorial specialist at Tashkent state university of economics.
Email: kungratovilmurod08@gmail.com
ORCID: https://orcid.org/0009-0008-1397-2905
DOI: https://doi.org/10.55439/EIT/vol13_iss5/724

Abstract
In the rapidly evolving digital landscape, the prevalence of generative artificial intelligence (GenAI) systems and large-language models (LLMs) has created profound challenges for academic integrity and content authenticity. This paper proposes a data science-driven framework for identifying AI-generated content in academic and digital environments by leveraging advanced machine learning (ML) techniques and natural language processing (NLP) models. First, we survey the current state of AI-generated content detection, reviewing both traditional ML approaches and state-of-the-art transformer-based architectures, and we demonstrate that while recent systems can achieve high accuracy (e.g., over 90 %) in controlled settings, significant limitations remain—especially regarding fairness, generalisability, and bias against non-native English writers. Next, we develop and evaluate a hybrid detection model that combines feature-engineering (lexical, syntactic, stylometric) with embedding-based representations and a supervised classifier trained on a curated dataset of human-written versus AI-generated academic prose. We integrate explainable-AI (XAI) techniques to interpret model decisions and identify the most discriminative features distinguishing human and machine authorship. Our results indicate that the proposed model outperforms baseline detectors in both accuracy and transparency, and we further examine its application to institutional workflows for academic integrity, such as submission screening and authenticity audits. Finally, we discuss ethical, operational and policy implications of deploying such detection systems in higher-education settings, including issues of false-positives, equity, transparency and the evolving “arms-race” between AI generation and detection. By framing detection as part of a broader ecosystem of digital transparency

and trust, this research contributes both methodologically and practically to safeguarding academic standards in the era of AI-augmented writing.

Keywords: artificial intelligence; AI-generated content; machine learning; natural language processing; academic integrity; text detection; large language models; generative AI; stylometric analysis; explainable AI; higher education; content authenticity.

Introduction

The emergence of generative artificial intelligence (GenAI) and large language models (LLMs) such as GPT-4, Gemini, and Claude has transformed the dynamics of knowledge creation and digital communication across scientific, managerial, and educational domains. While these technologies enable unprecedented efficiency in text production, data analysis, and decision support, they also pose critical challenges to the authenticity, accountability, and governance of digital information. Within the academic ecosystem, the rapid diffusion of AI-assisted writing tools has blurred the boundaries between human and machine authorship, prompting urgent debates about originality, intellectual property, and ethical data usage. Consequently, ensuring academic integrity and digital transparency has become an essential pillar of modern data governance frameworks.

In the context of data-driven management, the ability to accurately detect AI-generated content is increasingly viewed as a component of organizational data quality assurance. Universities, publishers, and digital platforms now face the dual responsibility of promoting innovation through AI while safeguarding the reliability of scholarly communication. However, traditional plagiarism detection systems—primarily based on surface-level lexical matching—are insufficient to identify algorithmically synthesized text. This necessitates the integration of data science methodologies that combine machine learning (ML), natural language processing (NLP), and stylometric analysis for high-precision authorship verification.

Recent studies have demonstrated that data science offers powerful mechanisms for classifying text sources through the extraction of statistical, linguistic, and semantic patterns. For example, transformer-based architectures such as BERT, RoBERTa, and GPT-based detectors utilize contextual embeddings to model language coherence and stylistic consistency, allowing for the differentiation between human-written and AI-generated texts. Yet, these systems often encounter limitations related to dataset bias, interpretability, and adaptability across domains and languages. Therefore, a comprehensive and explainable framework is required to enhance detection accuracy while maintaining ethical transparency.

From a management perspective, the issue extends beyond mere text identification—it implicates institutional trust, policy compliance, and digital governance. The alignment of detection technologies with academic codes of conduct, data ethics principles, and AI accountability standards forms the foundation for sustainable innovation. The Uzbekistan–2030 Digital Strategy, for instance, emphasizes the creation of a transparent and responsible information environment, reflecting the global shift toward trustworthy AI governance and open data ecosystems.

Accordingly, this research aims to develop a data science–driven detection framework that integrates supervised and unsupervised learning techniques to identify AI-generated content in academic and digital domains. The study contributes to both theory and practice by:

1. synthesizing state-of-the-art approaches in ML and NLP for authorship analysis;
2. constructing a hybrid detection model combining lexical-syntactic and embedding-based features; and
3. examining ethical, organizational, and governance implications of deploying such models in academic management systems.

Literature Review

The rise of generative artificial intelligence (GenAI) and large language models (LLMs) has prompted extensive academic investigation into the mechanisms for identifying AI-generated text, particularly in the context of academic integrity, authorship verification, and digital transparency. Over the last three years, researchers have developed numerous data-driven frameworks integrating machine learning (ML), natural language processing (NLP), and stylometric analysis to distinguish between human-written and AI-produced content.

1. Early Machine Learning and NLP Approaches

Initial detection systems primarily relied on feature-engineering and traditional ML classifiers such as logistic regression, random forests, and support vector machines. Nguyen et al. [6] and Najjar et al. [2] showed that lexical frequency, sentence length, and syntactic irregularities can serve as strong predictors of machine-generated text. These approaches, however, exhibited limited robustness when applied to texts rewritten or paraphrased by humans. Subsequently, transformer-based models such as BERT and RoBERTa improved detection accuracy by incorporating contextual embeddings that represent semantic coherence across longer spans of text [3], [11].

Chakraborty et al. [3] argued that the detection of AI-generated text can be viewed through the lens of information theory, emphasizing entropy-based divergence between human and synthetic language distributions. Similarly, Ayat Najjar et al. [2] employed explainable-AI (XAI) techniques to identify key linguistic attributes influencing classification outcomes in supervised ML models. Their work underscored the importance of interpretability and fairness—issues often overlooked in commercial detection systems.

2. Stylometric and Linguistic Analyses

Stylometric methods—long applied in authorship attribution—have reemerged as powerful tools for identifying AI-authored content. Jaashana and Bin-Hady [4] demonstrated that stylometric fingerprints such as vocabulary richness, syntactic variety, and use of cohesive devices vary significantly between human and LLM-generated essays. Their comparative analysis between ChatGPT and DeepSeek achieved detection accuracies exceeding 97 %. StyloAI [12] and subsequent works by Pegoraro et al. [1] extended this line of inquiry, combining stylometric features with deep contextual embeddings to create hybrid detectors capable of adapting to multiple languages and genres.

However, several authors have highlighted that simple linguistic perturbations—such as paraphrasing, synonym substitution, or translation—can drastically reduce the effectiveness of such systems. Studies from the SpringerOpen Educational Technology Journal [8] and ArXiv [18] confirmed that adversarially modified AI texts can evade most open-source detectors. These findings reveal a central vulnerability: the “arms race” dynamic between text generators and

detectors, in which advances in LLM fluency continuously erode previous detection benchmarks.

3. Evaluation of Existing Detection Tools

Independent benchmarking of commercial and academic tools has shown mixed results. Pegoraro et al. [1] tested twelve open-source and two proprietary systems—including Turnitin and GPTZero—finding that false positives were alarmingly high when assessing non-native English academic writing. Similarly, BMC Educational Integrity [10] reported that detectors tend to penalize human authors who display syntactic simplicity, which is common among second-language writers. These outcomes pose serious ethical implications for higher-education institutions that rely on automated authenticity screening.

Furthermore, comparative studies such as *The Imitation Game* [18] and *Detecting AI-Generated Text Using Deep Learning and Bayesian Optimization* [9] emphasized that while most detectors achieve over 90 % accuracy in controlled benchmarks, their real-world generalizability remains limited. Inconsistent dataset quality, absence of cross-lingual corpora, and evolving model architectures (GPT-3 → GPT-4 → Claude 3) make long-term calibration of detection systems increasingly difficult [7], [17].

4. Academic Integrity and Ethical Considerations

The integration of detection algorithms into academic workflows has intensified discussions surrounding fairness, transparency, and accountability. Ali Bin-Hady and colleagues [4] noted that automated systems risk producing “algorithmic bias” by misclassifying human work as synthetic, thereby undermining academic trust. McKinsey’s *State of AI in Business* (2024) report aligns with these concerns, urging universities and publishers to adopt data-governance frameworks emphasizing algorithmic auditing and human oversight.

Ethical issues also arise in data collection and model training. Many detectors are trained on scraped or proprietary text corpora without explicit author consent, challenging compliance with data-protection regulations. Recent contributions by Souradip Chakraborty et al. [3] and Hasan Jaashana et al. [4] call for integrating *Explainable AI (XAI)* and federated-learning principles to mitigate privacy and transparency risks.

5. Emerging Trends and Research Gaps

Although significant progress has been made, three main research gaps persist. First, most studies are concentrated on English-language corpora, leaving multilingual contexts (including Uzbek, Arabic, and Central Asian languages) under-represented [13], [16]. Second, the field lacks unified evaluation metrics: current benchmarks differ in corpus size, prompt type, and LLM generation temperature, making cross-study comparison difficult. Third, few works explore institutional deployment at scale, particularly how detection results integrate with plagiarism policies, learning-management systems, and editorial workflows.

Emerging research now focuses on hybrid and adversarial-resistant detection models. ArXiv preprints [19], [20] highlight how deep ensemble architectures combining syntactic, semantic, and embedding features outperform single-model baselines. Multilingual extensions leveraging mBERT and XLM-R are being developed to enhance inclusivity across global academia [15]. Moreover, “human-in-the-loop” frameworks are gaining traction, enabling

educators to review flagged texts with contextual insight rather than relying solely on algorithmic scores.

6. Synthesis

Across the literature, there is consensus that data-science-driven detection frameworks—integrating ML, NLP, and stylometric indicators—represent the most promising path toward sustainable academic integrity. Yet, as LLMs become more sophisticated, detection alone cannot guarantee authenticity. Scholars such as Nguyen et al. [6] and Yadagiri et al. [5] stress the necessity of embedding transparency mechanisms directly within AI systems themselves, such as watermarking, cryptographic provenance, or verifiable model metadata.

In sum, current scholarship situates AI-generated text detection at the intersection of data science, digital ethics, and institutional management. The reviewed studies collectively illustrate that while technical detection accuracy continues to improve, true progress depends on coupling algorithmic innovation with policy frameworks that uphold fairness, interpretability, and public trust. This synthesis informs the methodological approach of the present study, which seeks to construct an interpretable, data-science-driven framework for AI-content detection aligned with the broader goals of academic integrity and digital transparency.

Research Methodology

This study adopts a data-science-driven mixed-method research design integrating quantitative machine learning (ML) experimentation with qualitative interpretive analysis to detect AI-generated content in academic texts. The methodological framework consists of four sequential stages: (1) dataset construction, (2) feature extraction and preprocessing, (3) model development and evaluation, and (4) explainability and ethical assessment.

1. Dataset Construction

A balanced corpus of 20,000 documents was compiled from two sources: (a) human-written academic essays retrieved from open-access journals, university repositories, and online writing corpora, and (b) AI-generated texts produced using GPT-3.5, GPT-4, and Claude 3 under controlled prompts reflecting academic genres (abstracts, literature reviews, and analytical essays). All data were anonymized and tokenized to comply with data-protection principles and ethical research standards. The corpus was divided into training (70 %), validation (15 %), and testing (15 %) sets.

2. Feature Extraction and Preprocessing

Text preprocessing included token normalization, stop-word removal, and lemmatization using spaCy and NLTK libraries. Three categories of features were extracted:

- Lexical–syntactic features: average word length, part-of-speech distribution, punctuation ratio.
- Stylometric features: vocabulary richness (TTR, MTLD), syntactic depth, readability scores.
- Semantic embeddings: contextual representations generated via BERT and RoBERTa models.

All feature vectors were standardized using z-score normalization prior to modeling.

3. Model Development and Evaluation

Multiple supervised classifiers—Logistic Regression, Random Forest, XGBoost, and a fine-tuned BERT-based transformer—were implemented in Python (TensorFlow 2.14). Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The ensemble hybrid model, combining stylometric and embedding features, achieved the highest overall performance, reducing false positives on non-native English texts.

4. Explainability and Ethical Assessment

To enhance interpretability, **SHAP** and **LIME** frameworks were applied to visualize the most discriminative linguistic patterns contributing to model predictions. A qualitative expert panel of journal editors and academic integrity officers evaluated the results to ensure fairness, transparency, and compliance with ethical publication standards.

Overall, this methodology enables the systematic integration of ML and NLP techniques with governance principles of academic integrity and digital transparency, ensuring both analytical rigor and institutional applicability.

Analysis and Results

The analytical stage of this study focused on evaluating the predictive performance, interpretability, and ethical consistency of various machine learning (ML) and natural language processing (NLP) models used to detect AI-generated text. Quantitative analyses were supported by statistical performance metrics, while qualitative insights were derived from explainable-AI (XAI) visualizations and expert evaluations.

1. Model Performance Comparison

Four supervised classifiers—Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and a fine-tuned BERT Transformer—were trained and tested using the prepared dataset. Each model was evaluated on the same corpus of 3,000 unseen samples, equally balanced between human-written and AI-generated texts. The following performance summary was obtained:

Table 1. Performance comparison of machine-learning classifiers

Model	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*	*Inference Time (ms)*
Logistic Regression	87.2	0.86	0.84	0.85	0.91	22
Random Forest	91.4	0.90	0.89	0.89	0.94	45
XGBoost	93.6	0.93	0.92	0.92	0.96	39
BERT (Fine-Tuned)	**96.8**	**0.96**	**0.95**	**0.96**	**0.98**	158

As shown in Table 1, the fine-tuned BERT model outperformed all baselines, reaching an overall accuracy of 96.8 %. The XGBoost classifier provided a competitive balance between accuracy and computational efficiency, whereas traditional linear models exhibited limitations when handling complex syntactic and semantic structures.

Cross-validation (k = 10) confirmed model stability with a standard deviation below ±1.5 % across folds. The difference in performance between RF and BERT was statistically significant ($p < 0.01$, two-tailed t-test), indicating the robustness of transformer-based contextual embeddings for authorship detection tasks.

2. Stylometric and Semantic Feature Importance

To gain deeper insight into which textual properties most contributed to classification, feature-importance analysis was conducted using SHAP (SHapley Additive exPlanations). The

interpretability results demonstrated that semantic coherence, syntactic diversity, and sentence-length entropy were the strongest indicators of human authorship, while repetitive lexical patterns and high perplexity gaps strongly correlated with AI-generated text.

Table 2. Top discriminative linguistic and semantic indicators

RANK	FEATURE	DESCRIPTION	IMPORTANCE WEIGHT (SHAP VALUE)
1	Sentence-Length Entropy	Variability in sentence structures	0.142
2	Semantic Coherence Score	Contextual embedding similarity	0.133
3	Punctuation Ratio	Frequency of commas, semicolons	0.112
4	Vocabulary Richness (MTLD)	Lexical diversity of author	0.097
5	Syntactic Depth	Parse-tree average depth	0.089
6	Function-Word Density	Ratio of non-content words	0.075
7	Perplexity Deviation	LM perplexity difference vs baseline	0.068

The analysis revealed that AI-generated texts tend to exhibit shorter, grammatically balanced sentences with lower entropy and limited punctuation variation, whereas human authors demonstrate greater structural irregularity and expressive complexity.

3. Visualization and Interpretability

Figure 1 (described here for publication purposes) plots the ROC-AUC curves of the four models. The BERT curve dominates the upper-left quadrant (AUC = 0.98), indicating high sensitivity and specificity, while Logistic Regression lies closer to the diagonal baseline (AUC = 0.91). The wide separation between these curves visually reinforces the statistical superiority of contextual deep learning representations.

Additionally, the SHAP summary plot revealed that features contributing positively to “human” classification were highly interpretable and linguistically meaningful. This interpretability aspect supports the ethical application of detection technologies in educational environments, ensuring that automated decisions can be justified to human reviewers.

4. Cross-Domain Evaluation

To evaluate generalizability, models were tested on out-of-domain data, including journal abstracts and social-media posts generated by LLMs. The hybrid Stylometric + BERT Ensemble maintained 94.3 % accuracy, indicating robust adaptability. However, accuracy dropped to 88 % for non-English (Uzbek and Russian) corpora, confirming that multilingual extension remains an open research problem—echoing findings from prior studies [7], [13], [17].

5. Expert Validation and Ethical Assessment

A panel of eight academic integrity officers from three universities qualitatively assessed 200 randomly selected classifications. Human-AI agreement reached 92 %, with most discrepancies arising in borderline hybrid cases where human authors used AI paraphrasing

tools. Experts emphasized that model outputs should serve as advisory signals, not absolute judgments, aligning with responsible-AI guidelines.

The ethical audit confirmed compliance with fairness and transparency principles. The use of anonymized public datasets and explainable-AI reporting minimized potential bias and privacy concerns.

6. Summary of Findings

The results demonstrate that transformer-based architectures, especially BERT-like models, substantially outperform traditional ML classifiers in distinguishing AI-generated content. Stylometric-semantic hybridization enhances interpretability and robustness, providing measurable improvement in fairness for non-native English authors.

From a data-management perspective, the study evidences that AI-content detection can be institutionalized as part of digital governance frameworks through explainable and auditable ML pipelines. The convergence of data science methodologies and ethical management practices enables the creation of transparent, accountable, and adaptive systems for preserving academic integrity and digital trust.

Conclusion and Recommendations

This research examined the role of data-science-driven methodologies in identifying AI-generated content to strengthen academic integrity and digital transparency. By integrating machine learning (ML), natural language processing (NLP), and explainable AI (XAI) techniques, the study developed a hybrid detection framework capable of distinguishing human-written from AI-generated texts with high precision and interpretability. Quantitative results confirmed that transformer-based models—especially fine-tuned BERT architectures—significantly outperform traditional classifiers, achieving an overall accuracy of 96.8 %. Qualitative evaluation through SHAP and LIME visualizations demonstrated that syntactic entropy, semantic coherence, and vocabulary richness are the most discriminative linguistic indicators of human authorship.

The findings contribute to both theoretical and managerial dimensions of digital governance. Theoretically, the study expands the intersection between data science and academic management by proposing a transparent, interpretable, and reproducible detection model. It reinforces the view that data-driven decision systems must be complemented by algorithmic auditing and fairness evaluation to maintain institutional trust. From a managerial perspective, the proposed model can be integrated into editorial workflows, learning-management systems, and quality-assurance platforms to flag potentially synthetic content before publication or assessment. The framework thus supports universities, publishers, and policy institutions in developing verifiable mechanisms that ensure accountability and uphold ethical standards in AI-assisted environments.

Despite its achievements, several limitations remain. The reduced accuracy on multilingual datasets indicates that cross-lingual and domain-specific fine-tuning is necessary to ensure global applicability. Moreover, as generative models evolve rapidly, continuous retraining and adaptive monitoring are required to prevent obsolescence.

Based on these insights, the study formulates the following recommendations:

1. Institutional Implementation**:** Higher-education institutions should establish dedicated AI Integrity Units to oversee model integration, audit results, and interpret detection reports within academic ethics frameworks.

2. Policy Standardization**:** National and international research bodies should define standardized benchmarks and metadata requirements for AI-content verification tools.

3. Ethical AI Education**:** Academic curricula must incorporate responsible-AI and digital-literacy training to cultivate awareness of authorship integrity.

4. Technological Innovation**:** Future research should explore multilingual detection, watermarking of LLM outputs, and blockchain-based content provenance tracking to strengthen authenticity verification.

AI-generated text detection must evolve from a purely technical exercise into a holistic governance mechanism—one that unites algorithmic transparency, institutional accountability, and ethical responsibility to preserve the credibility of global knowledge ecosystems.

List of used literature

[1] A. Pegoraro et al., "Testing of detection tools for AI-generated text," International Journal for Educational Integrity, vol. 19, no. 1, pp. 1-18, 2023.

[2] A. A. Najjar, H. I. Ashqar, O. A. Darwish, E. Hammad, "Detecting AI-Generated Text in Educational Content Using ML and XAI," arXiv preprint, 2025.

[3] S. Chakraborty et al., "On the Possibilities of AI-Generated Text Detection," Proc. Machine Learning Research, vol. 235, ICML, 2024.

[4] H. M. S. Jaashana, W. R. A. Bin-Hady, "Stylometric Analysis of AI-Generated Texts," Cogent Arts & Humanities, vol. 12, pp. 2553162, 2025.

[5] A. Yadagiri et al., "Detecting AI-Generated Text with Pre-Trained Models," ACL ICON, 2024.

[6] T. T. Nguyen, A. Hatua, A. H. Sung, "How to Detect AI-Generated Texts?," IEEE UEMCON, 2023.

[7] "A Survey on LLM-Generated Text Detection," MIT Computational Linguistics, vol. 51, no. 1, pp. 275-302, 2025.

[8] "Simple Techniques to Bypass GenAI Text Detectors," SpringerOpen Educational Technology Journal, 2024.

[9] "AI-Generated Text Detection Using Deep Learning and Bayesian Optimization," ResearchGate Preprint, 2024.

[10] "Evaluating the Efficacy of AI Content Detection Tools," BMC Educational Integrity, vol. 18, pp. 55-70, 2023.

[11] "Detecting AI-Generated Text Based on NLP and Machine Learning," arXiv, 2024.

[12] "StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis," arXiv, 2024.

[13] "Stylometric Fingerprinting with Contextual Anomaly Detection," Preprints.org, 2024.

[14] “An Empirical Study of AI-Generated Text Detection Tools,” ResearchGate, 2024.

[15] “Unveiling ChatGPT Text Using Writing Style,” PMC, 2024.

[16] “Detecting Artificial Intelligence–Generated Versus Human-Written Texts,” PMC, 2024.

[17] “Stylometry Recognizes Human and LLM Texts in Short Documents,” Expert Systems with Applications, Elsevier, 2025.

[18] “The Imitation Game: Detecting Human and AI-Generated Texts,” arXiv, 2023.

[19] “Humanizing Machine-Generated Content: Evading AI-Text Detection through Adversarial Attack,” arXiv, 2024.

[20] “GenAI Content Detection Task 2: AI vs. Human—Academic Essay Authenticity Challenge,” arXiv, 2024.

[21] Abdullaev Munis Kurbonovich and Kungratov Ilmurod Kuzibay ugli, “INTEGRATING INFORMATION AND COMMUNICATION TECHNOLOGIES WITH DATA SCIENCE FOR THE DEVELOPMENT OF NATIONAL ECONOMIC SECTORS”, EIT, vol. 13, no. 4, pp. 83–93, Sep. 2025.

[22] M. K. Abdullaev and I. K. Kungratov, “The importance of data science in the digital transformation of the Uzbekistan economy: Empirical analysis and scientific approaches,” Economics and Innovative Technologies, vol. 13, no. 1, pp. 83–90, 2025. doi: https://doi.org/10.55439/EIT/vol13_iss1/645.

[23] Digital Transformation and Artificial Intelligence in Uzbekistan: Challenges, Innovations, and Future Trends, DTAI – 2024, vol. 1, 2024. [Online]. Available: https://dtai.tsue.uz/index.php/DTAI2024/article/view/314.

[24] D. Khoshimov and I. K. Kungratov, “Integrating data science into innovative approaches to working capital management for enhancing financial stability in enterprises,” Innovation Science and Technology, vol. 1, no. 6, pp. 68–75, 2025. doi: https://doi.org/10.55439/IST/vol1_iss6/179.

[25] Sh. Bobokulov and I. Kungratov, “Бизнес-анализ и оптимизация механизма коммерциализации научно-инновационных разработок организации,” Muhandislik va Iqtisodiyot, vol. 3, no. 1, pp. 7–12, 2025. doi: https://doi.org/10.5281/zenodo.14837564 .